UNIT –IV

Data exploration:

Exploration, attribute data query, spatial data query, raster data query, geographic visualization

Q.1. Explain the term 'Data Exploration' with its various aspects

Data Exploration

- Statistician have traditionally used variety of graphic techniques and descriptive statistics to examine data prior to more formal and structured data analysis.
- The windows operating system come up with multiple and dynamic link for windows, has further assisted exploratory data analysis by a allowing the user to directly manipulate data types in charts and diagrams.

Descriptive statistics

- Descriptive statistics summarizes the value of a dataset.
- They include the following

  o Range: the difference between the minimum and maximum values.

  o Median: the midpoint value, or the 50th percentile.

  o First quartile: the 25th percentile.

  o Third quartile: call the 75th percentile.

  o Mean: the average of data values. The mean can be calculated by Σxi/n,=1 where xi is the ith value and n is the number of values.

  o Variance: the average of the squared deviations of each data value about the mean. The variance can be calculated by

$$\sum_{i=1}^{n}(\text{xi} - \text{mean})2/\text{n}$$

o Standard deviation: the square root of the variance.

o Z score: a standardized score that can be computed by (x-mean)/s, where s is the standard deviation.

Graphs

Different types of graphs are used for data exploration. A line graph displays data as a line. The line graph example in figure 1 shows the rate of population change in the United States along the y-axis and the state along the x-axis. Notice a couple of "peaks" in the line graph.

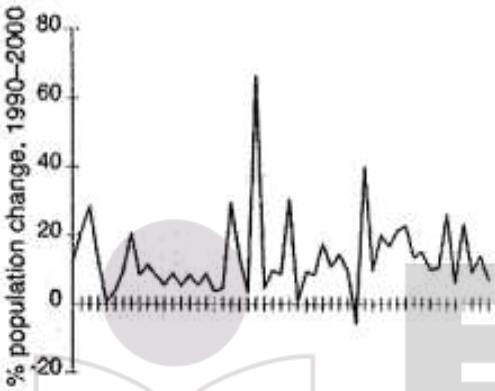


Fig. 1 : A line graph

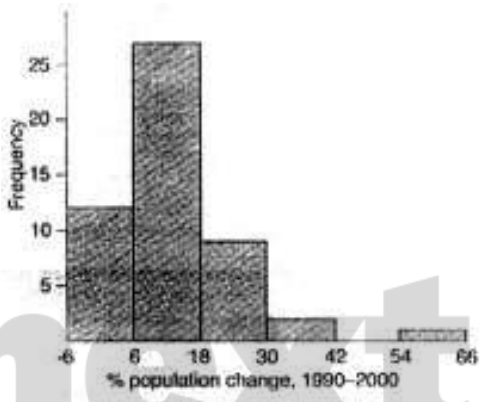
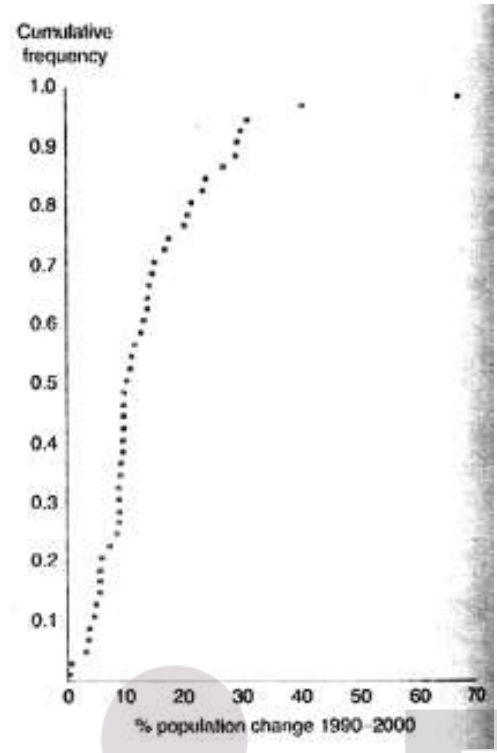


Fig. 2 : A histogram (bar chart)

A bar chart, also called a histogram, groups data into equal intervals and uses bars to show the number of frequency of values falling within each class. A bar chart may have vertical bars or horizontal bars. Figure 2 uses a vertical bar chart to group rates of population change in the United States into six classes. Notice one bar at the high end of the histogram.



A cumulative distribution graph is one type of line graph that plots the ordered data values against the cumulative distribution values. The cumulative distribution value of the ith ordered value is typically calculated as (i – 0.5)/n, where n is the number of values. This computational formula converts the values of a data set to within the range of 0.0 to 1.0. Figure 3 show a cumulative distribution graph.

Fig. 3 : A cumulative distribution graph

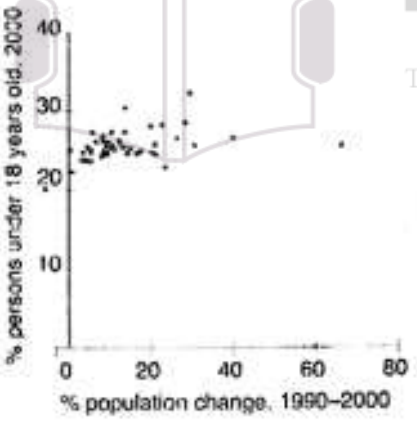


Fig. 4

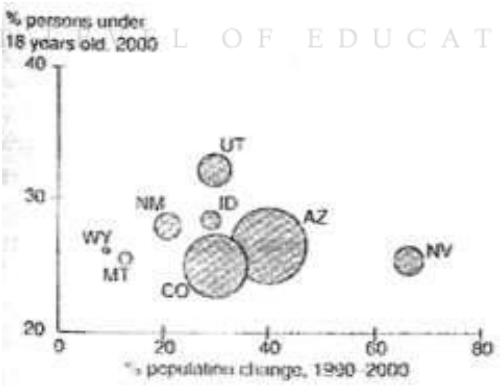


Fig. 5

Fig. 4 : A scatterplot plotting percent persons 18 years old in 2000 against percent population change, 1990-2000. A weak-positive relationship is present scatterplot uses markings to plot the values of two variables along the x– and y–axes. Figure 4 plots percent population change1990–2000 against percent persons under 18 years old in 2000 by state in the United States. The scatterplot suggests a weak positive relationship between the two variables.

Fig. 5 : A bubble plot showing percent population change 1990-2000, percent persons under 18 years old in 2000, and state population Bubble plots are a variation of scatterplots. Instead of using constant symbols as in a scatterplot, a bubble plot has varying sized bubbles that are

made proportional to the value of a third variable. Figure 5 is a variation of Figure 4, the additional variable shows by the bubble size is the state population ins 2000. As an illustration, Figure 5 only shown states in the Mountain region, one of the nine regions defined by the U.S. Census Bureau.

Boxplots, also called the "box and whisker" plots, summarized the distribution of five statistics from a data set the minimum, first quartile, median, third quartile, and maximum. By examining the position of the statistics in a boxplot, we can tell if the distribution of data values is symmetric or skewed and if there are unusually data points (i.e. outliers). Figure 6 shows a boxplot based on the rate of population change in the United States. This data set is clearly skewed toward the higher end. Figure 7 summarizes three basic types of data sets in terms of distribution of data values. Boxplots are therefore useful for comparisons between different data sets.

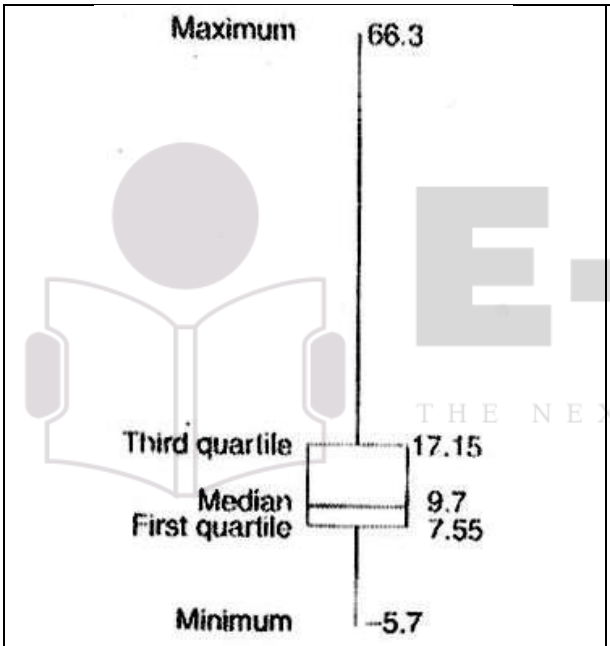


Fig. 6 : A boxplot based on the percent population change 1990-2000 data set.

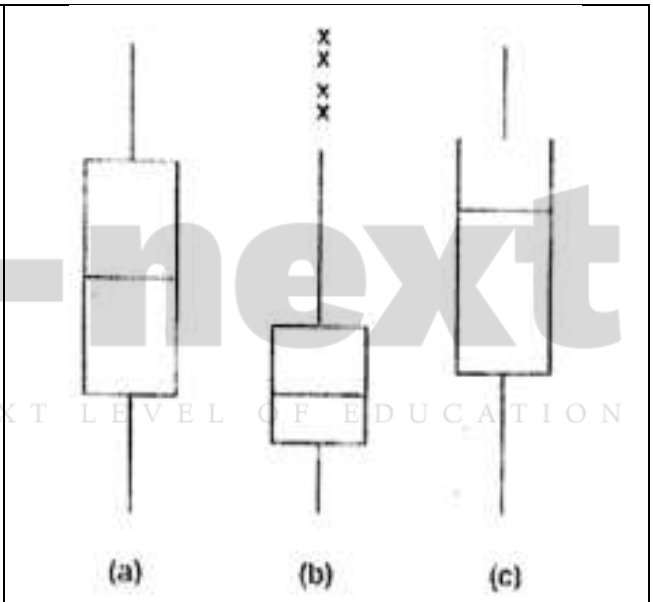


Fig. 7 : Boxplot (a) suggests that the data values follow a normal distribution. Boxplot (b) shows a positively skewed distribution with a higher concentration of data values near the high end. The x's in (b) may represent outliers, which are more than 1.5 box lengths from the end of the box. Boxplot (c) shows a

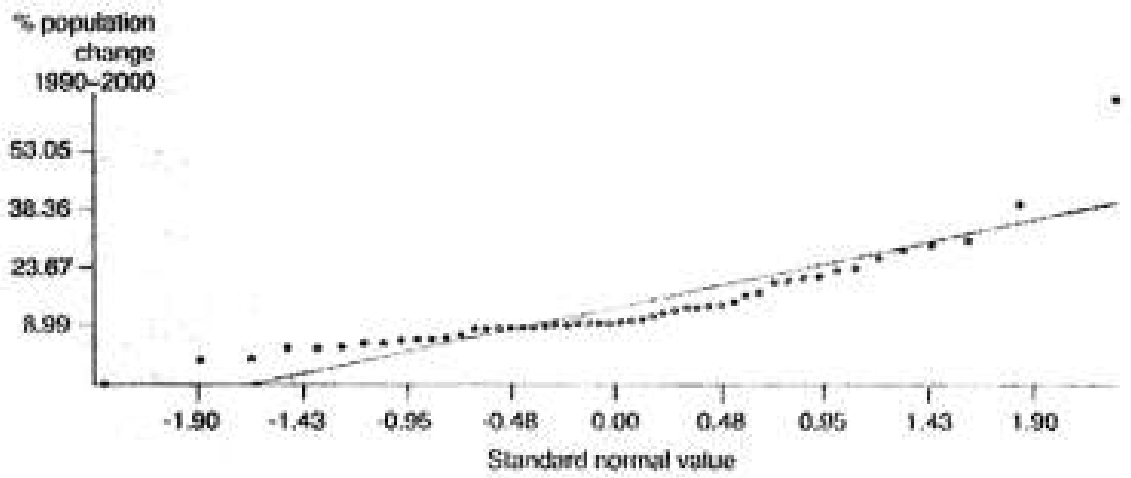


Fig. 8 : A QQ plot plotting percent population change, 1990-2000 against the standardized value from a normal distribution.

Some graphs are more specialized. Quantile-quantile plots, also called QQ plots, compare the cumulative distribution of a data set with that of some theoretical distribution such as the normal distribution, a bell-shaped frequency distribution.

The points in a QQ plot fall along a straight line if the data set follows the theorectical distribution. Figure 8 plots the rate of population change against the standardized value from a normal distribution. It shows that the data set is not normally distributed. The main departure occurs at the two highest values, which are also highlighted in previous graphs.

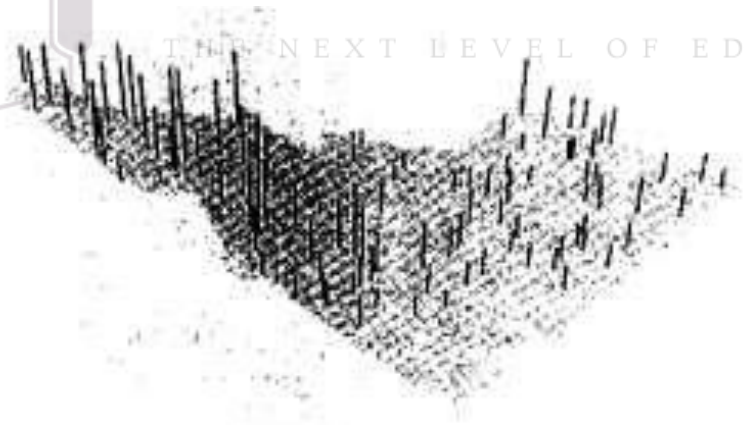


Fig.9 : A 3–D plot showing annual precipitation at 105 weather stations in Idaho. A north-tosouth decreasing trend is apparent in the plot.

Some graphs are designed for spatial data. Figure 9, for example, shows a plot of spatial data values by raising a bar at each point location so that the height of the bar is proportionate to its value. This kind of plot allows the user to see the general trends among the data values in both the x-dimension (east-west) and y-dimension (north-south).

Dynamic Graphics

When graphs are displayed in multiple and dynamically linked windows, they become dynamic graphs. We can directly manipulate data points in dynamic graphs. For example, we can pose a query in one windows and get the response in other windows, all in the same visual field. By viewing selected data points highlighted in multiple windows, we can hypothesize any patterns or relationships that may exist in the data. This is why multiple linked views have been described as the optimal framework for posing queries about data (Buja et al. 1996).

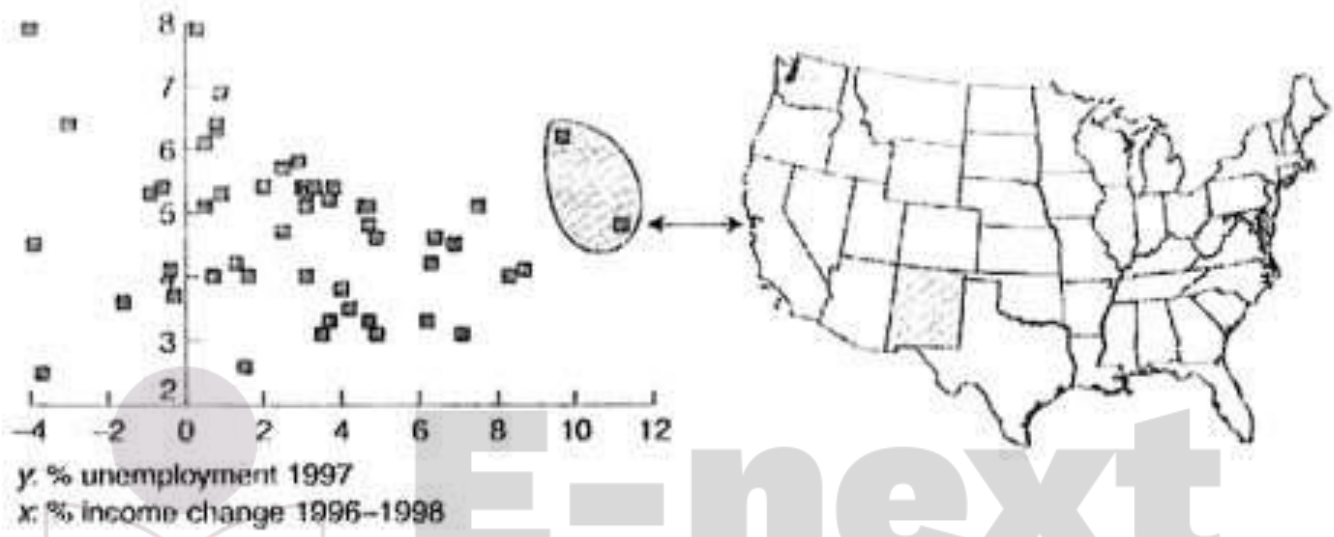


Fig. 10 : The scatterplot on the left is dynamically linked to the map on the right. The "brushing" of two data points in the scatterplot highlights the corresponding states (Washington and New Mexico) on the map.

A common method for manipulating dynamic graphs is brushing, which allows the user to graphically select a subset of points from a scatter plot and views related data points in other graphics (Backer and Cleveland, 1987). Brushing can be extended to maps (Monmonier 1989). Figure 10 illustrates a brushing example that links a scatter plot and a map. May GIS packages including ArcGIS have implemented brushing in the graphical user interface?

Q.2. Explain the term 'Attribute Data Query' with its various aspects.

(A) Attribute Data Query

- Attribute data query retrieve a data subset by working with attribute data.
- The selected data subset can be simultaneously examined in the table, displayed in chart and linked to the highlighted features in the map.
- The selected data subset can also be set for further processing.

SQL (Structured Query Language)

- SQL is a data query language design for relational databases.

- To use SQL to access database, we must follow the structure i.e. syntax of the query language
- The basic syntax of SQL
  select <attributes list >
  from <relation>
  where <condition>
- The select keyword selects field from the database, the from keyword selects tables from the database, and the where keyword specifies the condition or criteria for data query.
- We are considering the following Table A

PIN	Owner_name
P101	Wang
P101	Chang
P102	Smith
P102	Jones
P103	Costello
P104	Smith

Relation 1: Owner

PIN	Sale_date	Acres	Zone_code	Zoning
P101	1-10-98	1.0	1	Residential
P102	10-6-68	3.0	2	Commercial
P103	3-7-97	2.5	2	Commercial
P104	7-30-78	1.0	1	Residential

Relation 2: Parcel

- Let us take an example. Suppose we have to find the sale date of the parcel coded P101.:

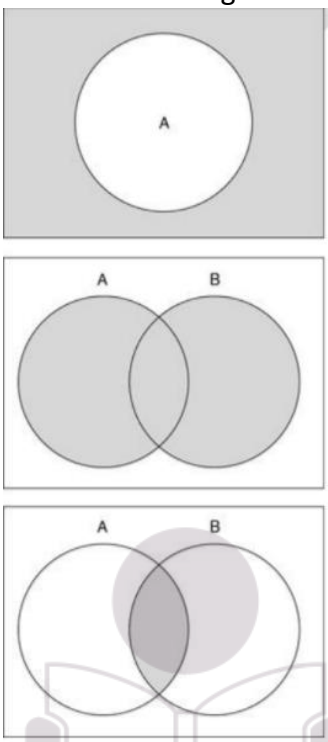
```
Select Parcel.Sale_Date
From Parcel
Where Parcel.Pin='P101'
```

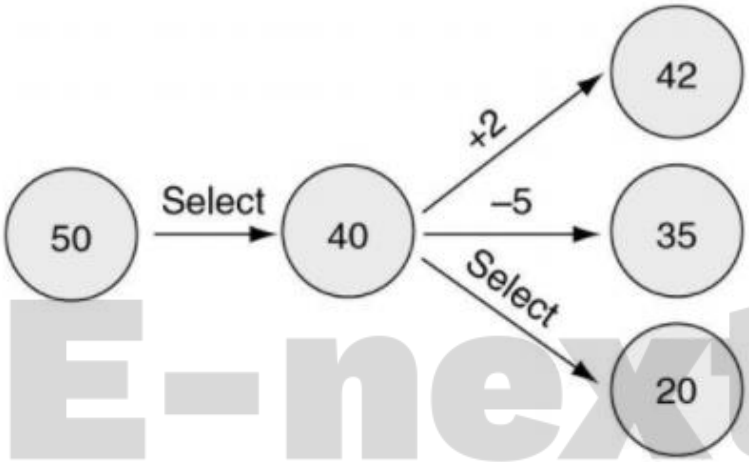
- The prefix for Parcel in Parcel.Sale_Date and Parcel. Pin indicates that the fields are from the parcel table.
- Suppose we have to find parcels that are greater than 2 acres and are zoned commercial :

```
Select Parcel.Pin
From Parcel
Where Parcel.Acres>2 And
Parcel.Zone_code=2
```

Type of operations:-

- Attribute data query begins with a complete data set.
- A basic query operation is to select a subset and divide the data set into 2 groups one containing selected records and the other unselected records.





- The different types of operations allow greater flexibility in data query for example, instead of using an expression of Parcel.Acres > 2 and Parcel.Zone_code =2, we can first use Parcel.Acres > 2 to select a subset and then use Parcel.Zone_Code =2 to select a subset from the previously selected subset. Examples of query operations

- **Example 1**

Q. Caller select a data subset and then add more records to it?

[Create a new selection] "cost" >=5 AND "soiltype"= 'Ns1'

Output: 0 of 10 records selected

[Add to current selection] "soiltype= 'N3'

Output: - 3 of 10 records selected.

- **Example 2:**

Q. Select a data subset and then switch selection?

[Create a new selection] "cost">8 OR "area>=400

Output: 2 of the 10 records selected.

[Switch selection]

Output: 8 of 10 records selected

- **Example 3 :**

Select a Data subset and then switch select a smaller subset from it?

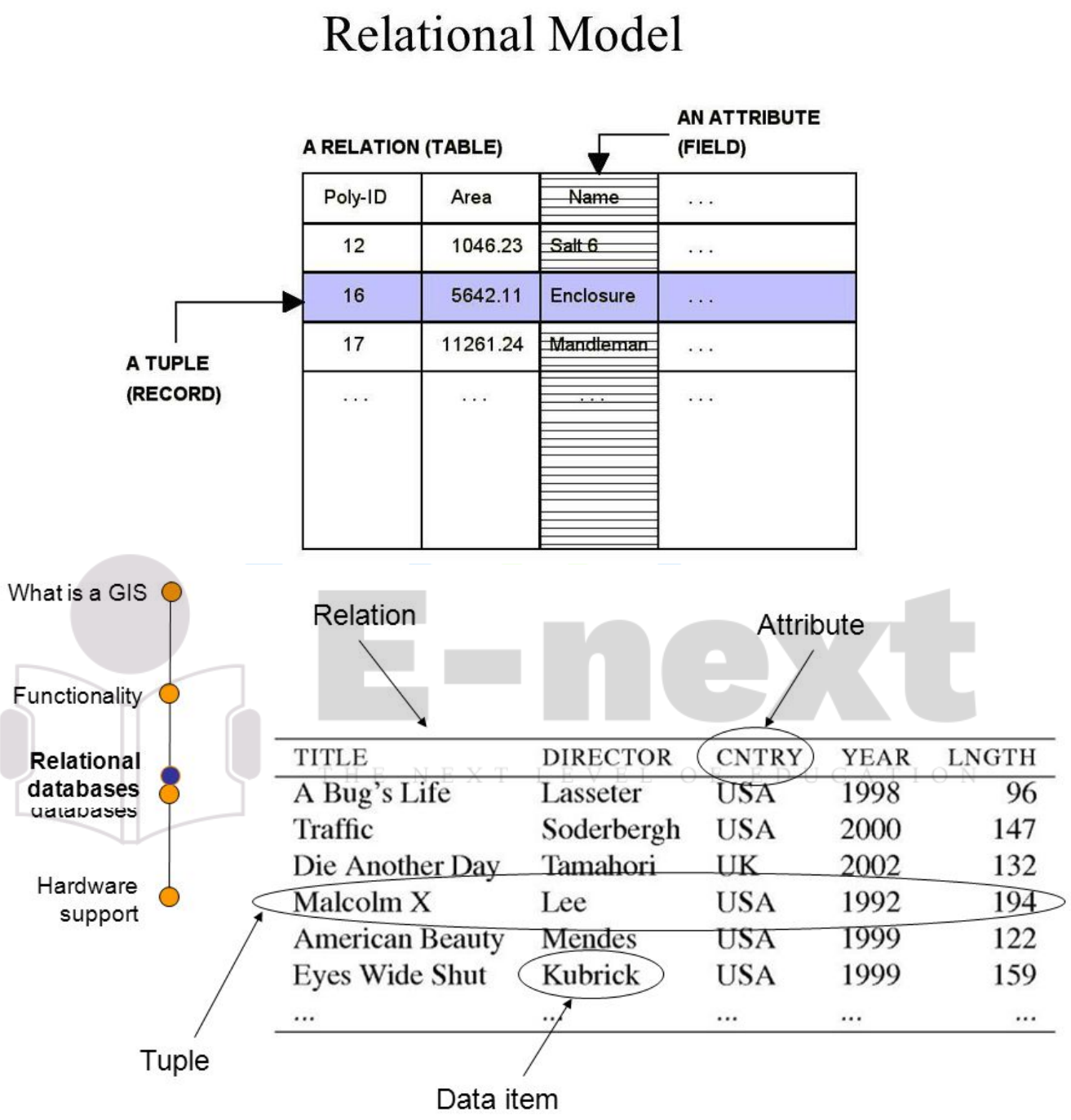
[Create a new selection] “cost” > 8 Or “area”>400

4 of 10 records selected

Q.3. Explain how an attribute query is executed on a relational GIS database.

Relational database query

- Relational database query works with a relational database, which may consist of many separate but interrelated tables.
- A query of a table in a relational database not only selects a data subset in the table but also selects records related to the subset in other tables.
- This feature is desirable in data exploration because it allows the user to examine related table characteristics from multiple tables.
- To use a relational database, we must be familiar with the overall structure of the database, the designation of keys in related tables, and a data dictionary listing and describing the fields in each table.
- For data query in 2 or more tables, we can choose to either join or relate the tables.
- A join operation combines attribute data from two or more tables into a single table.
- A relate operation dynamically links the tables but keeps the tables separate.
- A relate operation on the other hand can be used with all four types of relationship
- When a record in one table is selected the league will automatically select and highlight the corresponding record or records in the related tables.
- Join operation is appropriate for the one to one or many to one relationship but inappropriate for the one to many or many to many relationship.



Q.4. Explain the term 'Spatial Data Query' with all its aspects.

Spatial data query

- Spatial data query refers to the process of retrieving data subset from a layer by working directly with features.
- We may select features using cursor, a graphic, or the spatial relationship between features.

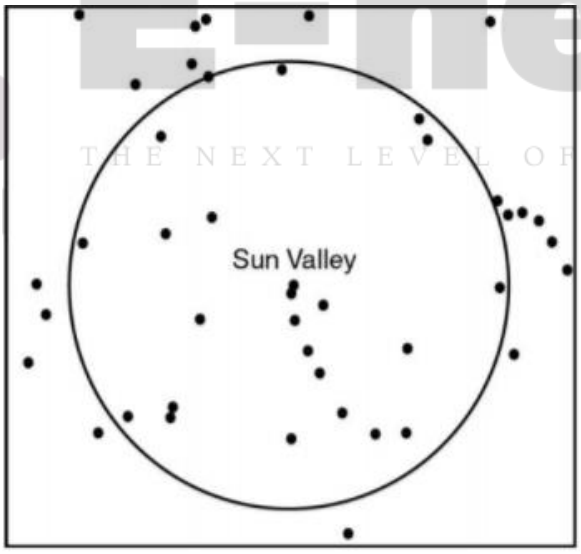
- As the geometric interface to the database, spatial data query complements attribute data query in data exploration.
- Similar to attribute data query, the results of spatial data query can be simultaneously inspected in the map, linked to the highlighted records in the table and displayed in charts.
- So they can also be saved as a new data set for further processing.
- Feature selection by cursor the simplest spatial data query is to select a feature by pointing at it or to select features by dragging a box around them.

Feature Selection by Cursor

- The simplest spatial data query is to select a feature by pointing at it or to select features by dragging a box around them.

Feature selection by graphic

- The query method uses a graphic such as a circle, a box, line or a polygon to select features the fall inside or what are intersected by the graphic object.
- We can draw the graphics for selection by using the mouse pointer.
- Example of query by graphic include selecting restaurants within 1 mile radius of a hotel, selecting land parcel that intersect a proposed highway.



Selection by spatial relationship

- This query method select feature based on their spatial relationships to other features.
- Features to be selected may be in the same layer as features for selection. Or more commonly they are in different layer.
- **Spatial relationships used for query include the following:**
  - **Containment** – selects features that fall completely within features for selection. Examples include finding schools within a selected county, and finding state parks within a selected state.

- **Intersect** – selects features that intersect features for selection. Examples include selecting land parcels that intersect a proposed road, and finding urban areas that intersect an active fault line.
- **Proximity**—selects features that are within a specified distance of features for selection. Examples include finding state parks within 10 miles of an interstate highway, and finding pet shops within 1 mile of selected streets. If features to be selected and features for selection share common boundaries and if the specified distance is 0, then proximity becomes adjacency. Examples of spatial adjacency include selecting land parcels that are adjacent to a flood zone, and finding vacant lots that are adjacent to a new theme park.

Combining Attribute and Spatial Data Queries :

So far we have approached data exploration through attribute data query or spatial data query. In many cases data exploration requires both types of queries. For example, both are needed to find gas stations that are within 1 mile of a freeway exit in southern. California and have an annual revenue exceeding $2 million each. Assuming that the layers of gas stations and freeway exits are available, there are at least two ways to answer the question.

1. Locate all freeway exits in the study area, and draw a circle around each exit with a 1 mile radius. Select gas stations within the circles through spatial data query. Then use attribute data query to find gas stations that have annual revenues exceeding $2 million.
2. Locate all gas stations in the study area, and select those stations with annual revenues exceeding $2 million through attribute data query. Next, use spatial data query to narrow the selection of gas stations to those within 1 mile of a freeway exit.

The first option queries spatial data and then attribute data. The process is reversed with the second option. Assuming that there are many more gas stations than freeway exits, the first option may be a better option, especially if the gas station map must be linked to other attribute tables for getting the revenue data.

The Combination Of spatial and attribute data queries opens wide the possibilities of data exploration. Some GIS users might even consider this kind of data exploration to be data analysis because that is what they need to do to solve most of their routine tasks.

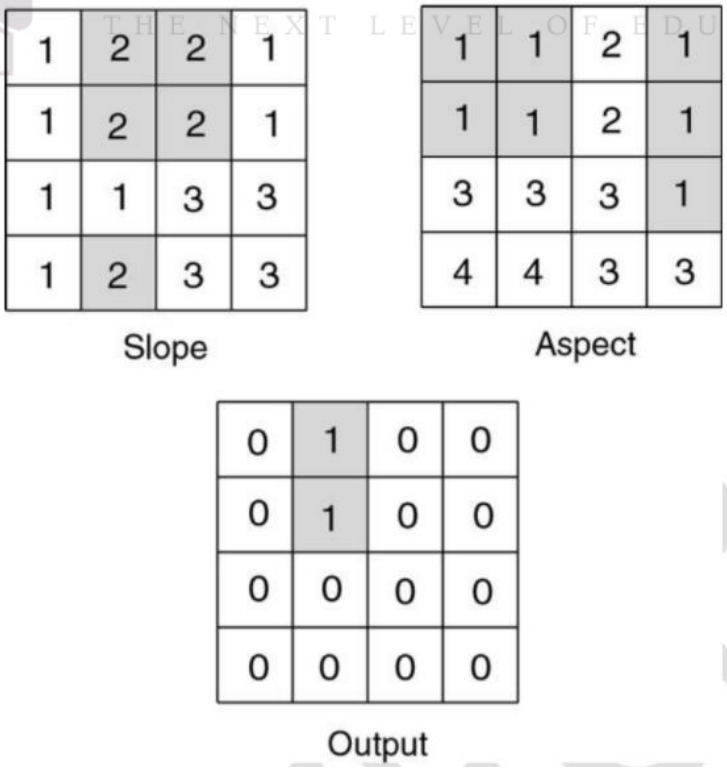
Q.5. Explain the term 'Raster Data Query' with all its aspects.

Raster Data Query

Although the concept and even some methods for data query are basically the same for both raster data and vector data, there are enough practical differences to warrant a separate section on raster data query.

Query by cell value

- The cell value in a raster typically represents a specific attribute value (example land use type, elevation, value etc.) at the cell location.
- Therefore the operand in raster data query is the raster itself rather than a field as in the case of vector data query.
- Raster data query uses a boolean statement separate cells that satisfy the query statement from cells that do not.
- The expression, [road=1] queries a road raster that has the cell value of 1.
- The operand [road] refers to the raster and the operand refers to a cell value, which may represent the interstate category.
- This next expression, [elevation]>1243.06 queries the floating point elevation raster that has the cell value greater than 1243.06.
- Because a floating point elevation raster contains continuous values, querying a specific value is not likely find any cell in the raster.
- Raster data query can also use the Boolean connectors of AND, OR and NOT to string together separate expressions.
- A compound statement with separate expressions usually applies to multiple rasters which may be integer or floating point or a mix of both types.
- For example, the statement, ([slope] =2) and ([aspect] =1), selects cells that have the value of 2 in the slope raster and 1 in the aspect raster.
- Those cells that satisfy the statement have the cell value of 1 on the output, while other cells have the cell value of 0.
- Figure l



Query by select feature:

- We can query a raster by using features such as points, circles, boxes or polygons.
- The query returns an output raster with value for cells that corresponds to the point locations or fall within the features for selection.

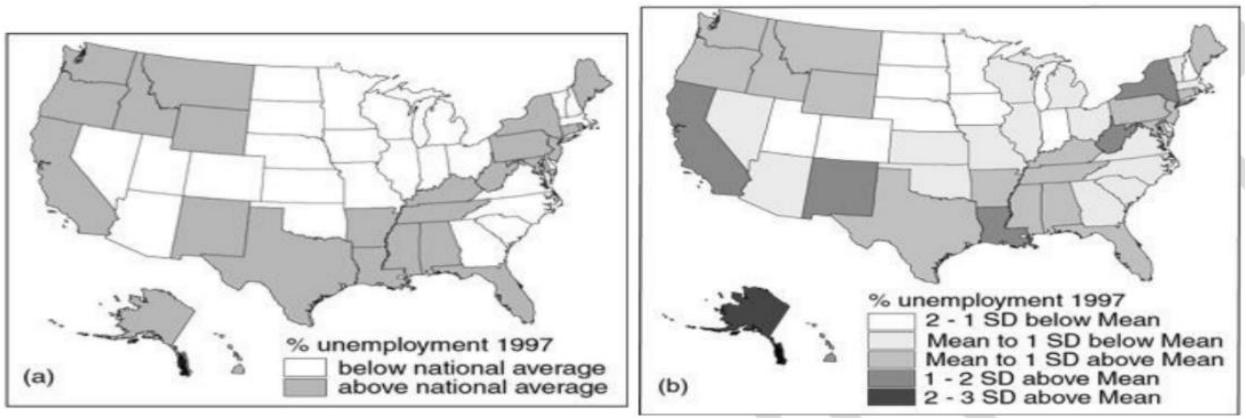
Q.6. Explain Geographic Visualization and its various techniques.

Map based data manipulation

- maps are an important part of GIS operations, including data exploration

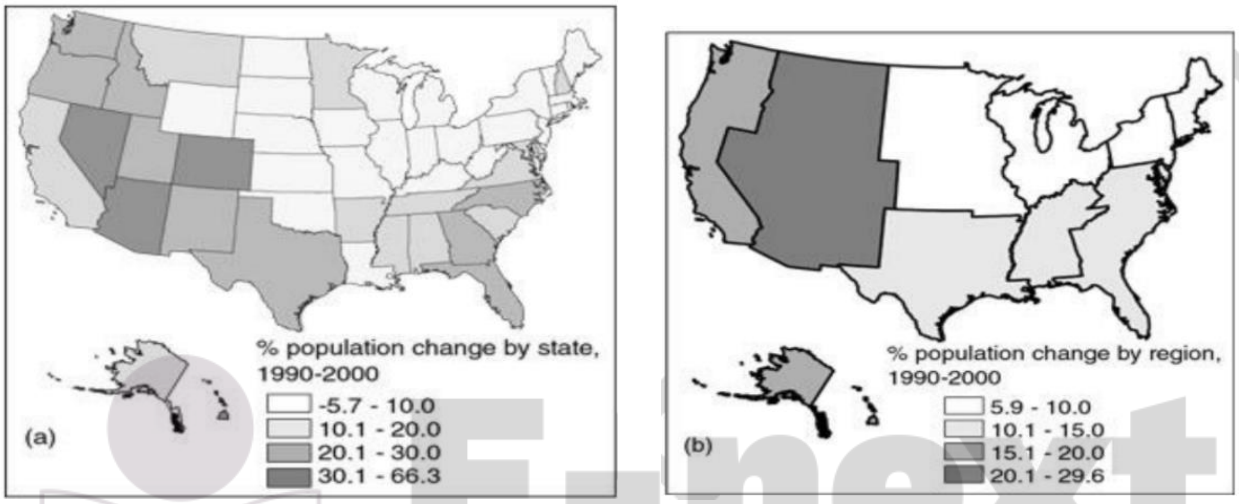
Data Classification

- Data classification can be a tool for data exploration, spatially if the classification is based on descriptive statistics.
- Suppose we want to explore rate of unemployment by state in the United States.
- To get a preliminary look at the data, we may place rate of unemployment into classes of above and below the national average. Figure J.a.
- Although generalized, the map divides the country into contiguous regions, which may suggest some regional factors for explaining employment.
- To isolate does states that are way above or below the national average, we can classify rate of unemployment by using the mean and standard deviation method figure J.b .
- We can now focus our attention on states that are, for example, more than one standard deviation above the mean.
- Classified maps can be linked with tables, graph and statistics for more data exploration activities. For example, we can link the maps in figure J with a table showing % change in median household income and find out whether states that have lower unemployment rates tend to have higher rates of income growth and vice versa.



Spatial aggregation

- Spatial aggregation is functional is similar to data classification except that it groups data spatially.
- Figure K shows % population change in the United States by state and by region.
- Used by the U.S. Census Bureau for data collection, regions are Spatial aggregate of states stop as shown in figure K.b , a map by region is a more general view of population growth in the country than a map by state does.



Map Comparison

- Map comparison can help a GIS user sort out the relationship between different maps. For example, the display of wildlife locations on a vegetation layer may reveal the association between the wildlife species and the distribution of vegetation covers.
- If the maps to be compared consist of only point or line features, they can be coded in different colors and superimposed on one another in a single view. But this process becomes difficult if they include polygon features or raster data. One option is to use transparency as a visual variable.
- A Semi transparent layer allows another layer to show through. For example, to compare two raster layers, we can display one layer in a color scheme and the other in semitransparent shades of gray.
- The gray shades simply darken the color symbols and do not produce confusing color mixtures. Another example is to use transparency for displaying temporal changes such as land cover change between 1990 and 2000. Because one layer is semitransparent, we can follow the areal extent of a land cover type from both years. But it is difficult to apply transparency to more than two layers.
- There are three other options for comparing polygon or raster layers. The first option is to place all polygon and raster layers, along with other point and line layers, onto the screen but to turn on and off polygon and raster layers so that only one of them is viewed at a time. Used by many websites for interactive mapping, this option is designed for casual users.